SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

July 2012 Vol.55 No.7: 618–625
doi: 10.1007/s11427-012-4343-6

A genome-wide association study of Alzheimer's disease using random forests and enrichment analysis

ZOU Liang, HUANG Qiong, LI Ao & WANG MingHui*

School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China

Received December 16, 2011; accepted June 5, 2012

Alzheimer's disease (AD) is a serious neurodegenerative disorder and its cause remains largely elusive. In past years, genome-wide association (GWA) studies have provided an effective means for AD research. However, the univariate method that is commonly used in GWA studies cannot effectively detect the biological mechanisms associated with this disease. In this study, we propose a new strategy for the GWA analysis of AD that combines random forests with enrichment analysis. First, backward feature selection using random forests was performed on a GWA dataset of AD patients carrying the apolipoprotein gene (APOEε4) and 1058 susceptible single nucleotide polymorphisms (SNPs) were detected, including several known AD-associated SNPs. Next, the susceptible SNPs were investigated by enrichment analysis and significantly-associated gene functional annotations, such as 'alternative splicing', 'glycoprotein', and 'neuron development', were successfully discovered, indicating that these biological mechanisms play important roles in the development of AD in APOEε4 carriers. These findings may provide insights into the pathogenesis of AD and helpful guidance for further studies. Furthermore, this strategy can easily be modified and applied to GWA studies of other complex diseases.

genome-wide association study, random forests, enrichment analysis, feature selection, Alzheimer's disease

Citation: Zou L, Huang Q, Li A, *et al*. A genome-wide association study of Alzheimer's disease using random forests and enrichment analysis. Sci China Life Sci, 2012, 55: 618−625, doi: 10.1007/s11427-012-4343-6

Alzheimer's disease (AD) is a common neurodegenerative disease that is characterized by degenerated memory, thinking and speaking impairment as well as disorders in behavior. Late-onset AD (LOAD) which usually occurs after 65 years of age accounts for 90% of AD patients [1]. As populations age, it has been estimated that the number of people affected by this disease will soar to about 80 million by 2040; this number includes 60 million people living in the developing countries, including China [2]. AD is already becoming a critical social and medical issue and further research into this debilitating disease is urgently required.

Although the pathogenesis of AD has not been elucidated completely, previous researches have demonstrated that genetic factors play a major role [3]. Recently, APOE (apolipoprotein E), NGFR (nerve growth factor receptor), and presenilin-1/presenilin-2 have been implicated in AD [4–8]. APOEε4 has been verified as a genetic risk factor for AD [6–8], and the abnormal expression of APOEε4 has been reported to lead to impairment of the cholinergic neuron and the abnormal increase of amyloidβ protein, which in turn triggers the formation of neurofibrillary tangles and disruption of the neuronal cytoskeleton [7]. Almost two-thirds of AD patients are APOEε4 carriers and the risk has been found to increase significantly with the number of APOE ε4 copies [8]. The associated mechanism is reported to be very complicated; for example, there are multiple susceptible genes, such as GAB2, that contribute to AD cooperatively with APOEε4 [9,10].

With the rapid development of high-throughput genotyping chips, genome-wide association (GWA) studies have

*Corresponding author (email: mhwang@ustc.edu.cn)

been widely employed in many researches on a variety of complex genetic diseases [9,11–18]. Under the "common disease, common variant" hypothesis, GWA studies have been used to screen thousands or millions of single nucleotide polymorphisms (SNPs) by comparing case and control subjects to detect disease-associated alleles [19]. Many GWA studies on AD [9,16,17] have been carried out in recent years and Reiman *et al* [9], for example, identified six susceptible SNPs in the GAB2 gene that were associated with AD.

Univariate methods, such as the chi-square test [9], the trend test [18] and the likelihood ratio test [20], are most commonly used in GWA studies. However, these tests assume mutual independency between features and, therefore, possible feature interactions are ignored [15,21]. Consequently, when applied to GAW studies, they may fail to detect SNPs that are associated with a disease via genetic interactions. Furthermore, it is unfeasible to use these methods to evaluate the statistical significance of some of the high-level biological mechanisms that have been implicated in complex diseases and which may better elucidate the pathogenesis of the disease. To address this problem, we proposed a novel strategy that combines random forests and enrichment analysis to analyze the GWA data. Our strategy is based on the following two premises: (i) random forests can delineate interactions between SNPs, and therefore can select disease-associated SNPs more effectively [15,21]; and (ii) enrichment analysis of the selected SNPs can detect significant functional annotations and corresponding molecular biology mechanisms [22,23]. We applied this strategy to a GWA dataset and discovered several biological mechanisms in APOEε4 carriers that were significantly associated with AD. The findings from this study may help further elucidate the pathogenesis of AD.

1 Materials and methods

1.1 Dataset

The GWA data of 644 APOEε4 carriers (117 cases and 527 controls, all over the age of 65) from a previous study on LOAD [9] was used in this work. After removing unqualified SNPs as described previously [9] (such as minor allele frequencies<2%, or Hardy Weinberg equilibrium *P*-value <0.01), the GWA data included 312316 SNPs from the total of 502267 SNPs that were genotyped by SNP-arrays.

1.2 Random forests

Random forest is a powerful statistical machine learning method that combines many independent decision trees [24,25]. To build a single decision tree, two stochastic processes are adopted. Firstly, bagging is used to construct a training subset by randomly sampling from the original data. Secondly, at each node of a decision tree, a subset of variables numbering m is selected randomly from all the p variables without replacement, and the optimal split based on this subset is used to split the node. These two random processes effectively reduce the correlation between independent trees, thereby improving the robustness of random forests and avoiding an over-fitting problem. Next, for tree t, the corresponding out-of-bag data is used as the test data for evaluating misclassification error. One commonly adopted formula of the importance of a feature variable v is defined as follows:

$$\text{importance}_v = \frac{1}{N}\sum_{t=1}^{N}(\text{count}_{t,ini} - \text{count}_{t,v}), \quad (1)$$

where N is the number of trees in the random forests, $count_{t,ini}$ is the number of correctly classified samples and $count_{t,v}$ is the number of correctly classified samples by tree t in a new test data in which v is randomly permuted. The importance parameter reflects the contribution of a feature variable to data classification. In addition, the hierarchical topology of decision trees is advantageous in handling complicated interactions between features; therefore, random forests have recently been used successfully in GWA studies [12–15].

In this study, we constructed random forests using Willows software [26] and performed backward feature selection to find susceptible SNPs in AD patients carrying APOEε4. We built the random forests using all available SNPs (denoted by S_0), and kept only the SNPs with positive importance (denoted by S_1). This procedure was repeated until all the remaining SNPs had positive importance. This optimal subset of SNPs was then used for enrichment analysis.

1.3 Enrichment analysis

By taking advantage of various bioinformatics databases and analysis tools, enrichment analysis aims to detect the correlation between a cohort of gene sets and a reference gene set that shares the same or similar functions and consequently discovers statistically significant biological mechanisms [22,23,27–31]. Statistical methods such as the Fisher test [22,23,30] and the chi-square test [31], are usually adopted in enrichment analysis. The DAVID software [22,23] that consists of integrated online databases and analytic tools, was employed in this study. For example, the "Gene Name Batch Viewer" can quickly map a list of SNPs to the corresponding genes. DAVID uses EASE scores [32] based on a modified Fisher test to measure the association between functional annotation and genes. The software also provides different approaches for detailed enrichment analysis, including a gene annotation chart and functional annotation clustering, which assisted our research on the AD-associated biological mechanisms in APOEε4 carriers.

2 Results and discussion

2.1 SNP screening based on random forests

Before selecting the AD-associated SNPs in APOEε4 carriers using random forests, the number of trees (N) and the number of variables (m) were determined based on the characteristics of the GWA data. The dimensionality of the data used in GWA studies is very high and the number of SNPs largely exceeds the sample size n. This phenomenon is known as the 'large p and small n' problem. A previous study [33] had shown that in such a situation, the m and N parameters should be large enough during the procedure of building random forests. Therefore, taking this requirement and computational efficiency into consideration, we set m equal to the square root (p) and N to greater than $2p$ in this study. Figure 1A illustrates the relationship of the number of selected SNPs and the number of times that feature selection was performed. Two stages were observed during this procedure. The first stage appeared at the beginning of feature selection, in which a large number of irrelevant SNPs were removed quickly. At the end of the first stage, the speed of feature selection decreased gradually. The subsequent second stage had a similar trend as the first except that feature selection became even slower at the end of this stage. Finally, 1058 AD-associated candidate SNPs were selected. A retrospection analysis of the overall change of the importance of these SNPs showed an obvious increasing trend of importance as feature selection was performed (Figure 1B). Specifically, in the first stage of feature selection, the overall importance increased slowly probably because of the existence of a large number of irrelevant features in the early stage of the process. However, after most of the disease-unrelated SNPs were removed, the signal-to-noise ratio of the data improved significantly in the second stage and the importance of the SNPs correspondingly increased. Finally it should be pointed out that the small local fluctuations on the curve in Figure 1B are probably caused by the randomness in the procedure that was used to construct the training/test data and decision trees.

To verify the results obtained by random forests, we analyzed the selected SNPs and their corresponding importance. Firstly, we calculated the P-values of all the SNPs using the chi-square test under the assumption that there was no association between these SNPs and AD, and compared the values with the importance of the corresponding SNPs (Figure 2). For a better illustration, we used local regression (loess) to generate the fitting curve. Figure 2 shows that there was an overall declining trend of importance with increases in the P-values. In addition, in the right part of the fitting curve, the SNPs' importance decreased significantly with increasing P-values while, in the left part of the fitting curve, this trend was not apparent. One major reason for this difference is that these SNPs cooperate together to affect the pathogenesis of the disease and therefore, using a univariate method, it is difficult to associate them with AD.

Next, we examined the distribution of the selected SNPs on the human genome and the results are shown in Figure 3A. We found that 96 of the SNPs were located on chromosome 11 which was significantly higher than the number of SNPs on the other chromosomes. An in-depth analysis of the SNP locations on chromosome 11 (Figure 3B) identified several SNPs with high importance at position 11q14.1, indicating this region of the chromosome may harbor susceptible genes in APOEε4 carriers. A literature search revealed that the GAB2 gene, which is located in this region, has a significant influence on the prevalence of AD [9,34,35]. In these earlier studies it was shown that GAB2 is associated with the phosphorylation of the tau protein and the formation of neurofibrillary tangles that play an important role in AD development. We analyzed all the

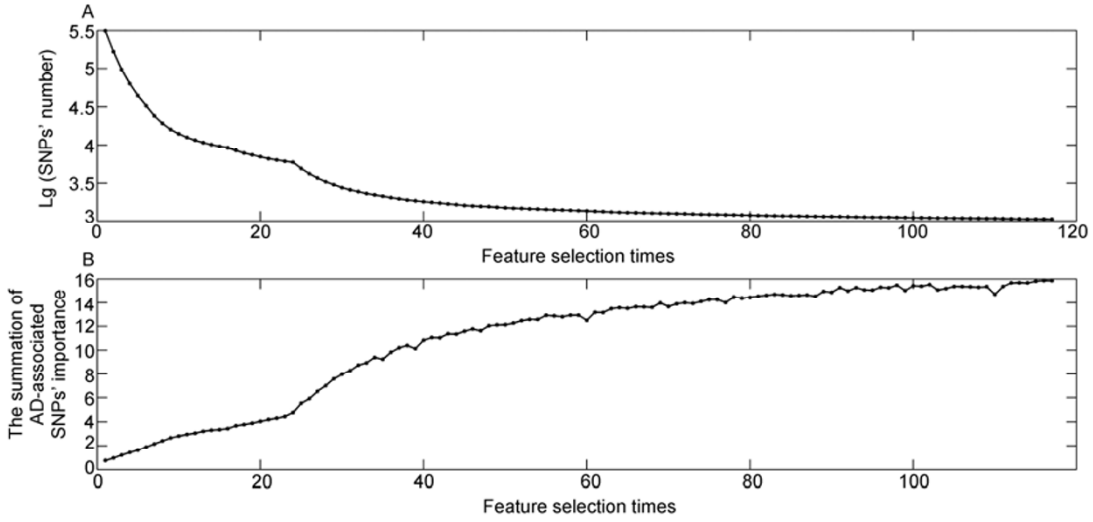

Figure 1 The feature selection procedure based on random forests for AD-associated SNPs. A, The number of SNPs in each feature selection run. B, The overall importance of the 1058 selected SNPs.

GAB2-related SNPs in our candidate SNP list and discovered that many of the SNPs were known to be associated with AD (Table 1). In addition, we found some AD-associated candidate genes, including APOC1 [36] and NGFR [4], in the same region of chromosome 11. APOC1 together with APOE has been demonstrated to influence the development of AD [36]. A SNP (rs4420638) in the APOC1 gene has also been verified to be significantly associated with AD [37].

2.2 Analysis of AD-associated biological mechanisms

We performed an enrichment analysis of the 1058 AD-associated candidate SNPs identified by random forests. Using the analytic tools provided in DAVID, 749 SNPs were mapped to 520 genetic functional annotations from various bioinformatics databases, such as Uniprot, PIR, GO, and InterPro; 144 of the annotations had *P*-value less than 0.01. To address the issue of multiple comparisons, the Benjamini correction was adopted and a threshold of 0.01 for the corrected *P*-values was used to determine statistical significance. Finally, 27 functional annotations were identified to be significantly associated with AD in APOEε4 carriers (Table 2).

The most significant functional annotation was "alternative splicing" with a corrected *P*-value of 1.13×10^{-8}, indi-

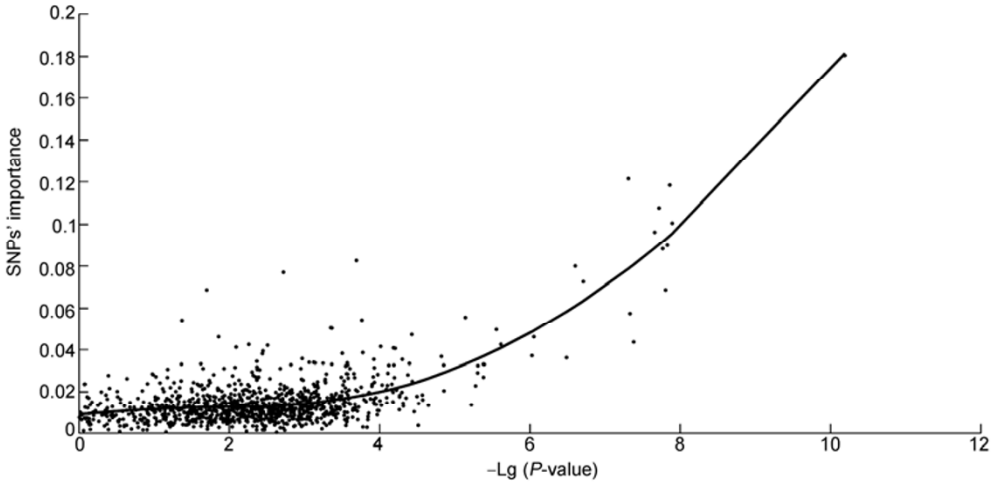

Figure 2 Comparison of the importance of the SNPs with the *P*-value calculated using the chi-square test (the fitted curve of importance against *P*-value was obtained by local regression).

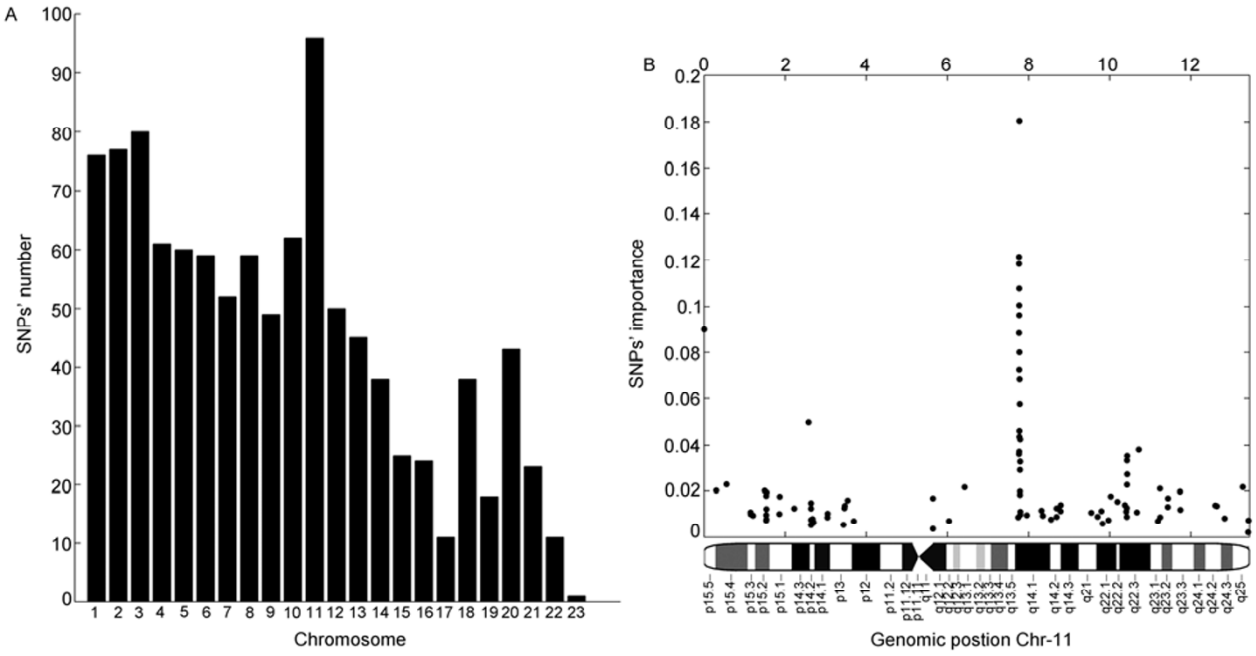

Figure 3 Distribution of the AD-associated SNPs on the human chromosomes. A, The distribution of the AD-associated candidate SNPs in APOEε4 carriers. B, The importance of AD-associated SNPs on chromosome 11 in APOEε4 carriers.

Table 1 Known susceptible SNPs in the GAB2 gene that are among the selected candidate SNPs

SNP_ID	Chromosome	Allele	Position
rs1385600	11	C/T	77613814
rs4945261	11	A/G	77667908
rs7101429	11	A/G	77670615
rs10793294	11	A/C	77674051
rs7115850	11	C/G	77722719
rs2373115	11	G/T	77768798

Table 2 Functional annotations of the AD-associated SNPs in APOEε4 carriers

Category	Functional annotation	*P*-value	Benjamini *P*-value
SP_PIR_KEYWORDS	Alternative splicing	2.40×10^{-11}	1.13×10^{-8}
UP_SEQ_FEATURE	Splice variant	7.33×10^{-11}	1.69×10^{-7}
SP_PIR_KEYWORDS	Glycoprotein	1.83×10^{-8}	4.31×10^{-6}
GOTERM_BP_FAT	Neuron development	1.04×10^{-8}	6.66×10^{-6}
GOTERM_BP_FAT	Neuron projection morphogenesis	1.39×10^{-8}	7.17×10^{-6}
GOTERM_BP_FAT	Cell morphogenesis involved in neuron differentiation	8.92×10^{-9}	7.64×10^{-6}
GOTERM_BP_FAT	Neuron projection development	6.97×10^{-9}	8.96×10^{-6}
GOTERM_BP_FAT	Cell morphogenesis involved in differentiation	2.64×10^{-8}	1.13×10^{-5}
GOTERM_BP_FAT	Neuron differentiation	3.18×10^{-8}	1.17×10^{-5}
GOTERM_BP_FAT	Axonogenesis	5.49×10^{-9}	1.41×10^{-5}
UP_SEQ_FEATURE	Glycosylation site: N-linked (GlcNAc...)	1.41×10^{-8}	1.62×10^{-5}
GOTERM_BP_FAT	Axon guidance	1.68×10^{-7}	5.41×10^{-5}
GOTERM_BP_FAT	Cell projection morphogenesis	3.34×10^{-7}	9.54×10^{-5}
INTERPRO	IPR013098: Immunoglobulin I-set	1.80×10^{-7}	2.15×10^{-4}
GOTERM_BP_FAT	Cell part morphogenesis	8.58×10^{-7}	2.21×10^{-4}
GOTERM_BP_FAT	Cell motion	2.13×10^{-6}	4.56×10^{-4}
GOTERM_BP_FAT	Cell projection organization	2.08×10^{-6}	4.86×10^{-4}
GOTERM_BP_FAT	Cell morphogenesis	6.47×10^{-6}	1.28×10^{-3}
UP_SEQ_FEATURE	Domain: Fibronectin type-III 1	3.21×10^{-6}	2.47×10^{-3}
UP_SEQ_FEATURE	Topological domain:Cytoplasmic	5.37×10^{-6}	2.47×10^{-3}
UP_SEQ_FEATURE	Domain: Ig-like C2-type 3	4.53×10^{-6}	2.61×10^{-3}
GOTERM_BP_FAT	Biological adhesion	1.75×10^{-5}	3.00×10^{-3}
GOTERM_BP_FAT	Cell adhesion	1.73×10^{-5}	3.18×10^{-3}
UP_SEQ_FEATURE	Domain: Fibronectin type-III 2	1.12×10^{-5}	4.30×10^{-3}
INTERPRO	IPR008957: Fibronectin, type III-like fold	7.33×10^{-6}	4.38×10^{-3}
GOTERM_BP_FAT	Cell-cell adhesion	3.08×10^{-5}	4.93×10^{-3}
GOTERM_BP_FAT	Cellular component morphogenesis	6.49×10^{-5}	9.77×10^{-3}

cating a strong association between alternative splicing and AD in the APOEε4 carriers. Alternative splicing is an important gene regulation mechanism, which promotes complexity of gene expression and diversity of proteins in eukaryotes [38]. Studies have shown that there is an intrinsic relationship between alternative splicing and the pathogenesis of AD [39,40]. For example, Tollervey *et al*. [39] investigated alternative splicing in patients with AD and other neurodegenerative diseases and identified two kinds of alternative splicing: one was age-related splicing changes in normal individuals and the other was disease-specific splicing changes that played an important role in regulating the genes involved in metabolism and DNA repairing in AD patients. Another significant functional annotation was "glycoprotein" which refers to proteins with oligosaccharide chains that are covalently attached to the polypeptide side chains. Glycoproteins usually participate in various physiological functions such as material transportation, nerve conduction, cell development and differentiation. It has been reported that excessive glycosylation of the tau protein triggers the instability of microtubules leading to serious degeneration or even disability of the neurofibrils which contributes to the emergence and development of AD [41]. Other significant AD-associated annotations have been reported in APOEε4 carriers [42,43], including "cell morphogenesis involved in neuron differentiation" and "neuron development". For comparison, we generated a list of significant SNPs detected by the univariate method; however, enrichment analysis failed to give any significantly associated functional annotations. One possible explanation for this result is that the univariate method does not take into account the synergy of multiple risk-factors and some

AD-associated SNPs that affect the disease indirectly were neglected, leading to the low abundance of functional annotations.

From the results of the enrichment analysis, we noticed that some seemingly different annotations actually represented similar or even the same functions. For example, the functional annotations "cell morphogenesis involved in neuron differentiation" and "cell morphogenesis involved in differentiation" have different names and *P*-values (Table 2), but both terms describe cell morphogenesis in differentiation. In the GO database the former term is a child node of the latter with the "is a" relationship, indicating that "cell morphogenesis involved in neuron differentiation" is a subtype of "cell morphogenesis involved in differentiation". To avoid redundant analysis and to better understand the relations between the biology mechanisms, we performed functional annotation clustering using DAVID. The most significant cluster included 13 different functional annotation terms and the frequency of occurrence of each of these terms in the cluster is shown in Figure 4. The term "axon guidance" was the least observed with an occurrence frequency of 0.27; the other annotations had an occurrence frequency more than 0.40 with "neuron differentiation" being the most frequently observed (0.65).

Because all the annotations in this cluster were from the GO database, we further investigated the semantic relationships between them. We listed all the GO annotations related to the "axon guidance" annotation (GO relationships are represented as either "is a" or "part of") and found that all the other annotations in the cluster were in fact parent nodes of "axon guidance", as illustrated in Figure 5. Specifically, the "axonogenesis" term was the only direct parent node and all the other annotation terms were indirect parent nodes. Interestingly, the parent nodes themselves were also highly related (Figure 5), indicating that the topology of the semantic relationships was consistent with the results of the functional annotation clustering. Finally, we performed a topological analysis by combing the *P*-values from the enrichment analysis, and found that the most significant annotation "neuron development" appeared in the middle of the map. More interestingly, we observed that the statistical significance of the annotations tended to decrease as the distance to "neuron development" in the map became longer. For example, with their increasing distance to "neuron development", the *P*-values of "cell morphogenesis involved in neuron differentiation" (7.6×10^{-6}), "axonogenesis" (1.4×10^{-5}) and "axon guidance" (5.4×10^{-5}) increased and became less significant. These findings suggest that neuron development is the most significant AD-associated biological mechanism in the APOEε4 carriers.

3 Conclusion

In this work, we introduced a new approach for the GWA study of AD in APOEε4 carriers. To detect disease-associated functional annotations and corresponding biological mechanisms, we proposed a data analysis strategy that combines random forests with enrichment analysis and found that significant AD-associated biological mechanisms, such as alternative splicing, glycoprotein, and neuron development, were successfully identified using this novel approach. These mechanisms have been reported previously in other studies, confirming the effectiveness of our method. Further studies on these biological mechanisms will help to better understand the pathogenesis of AD and provide guidance for clinical treatment plans.

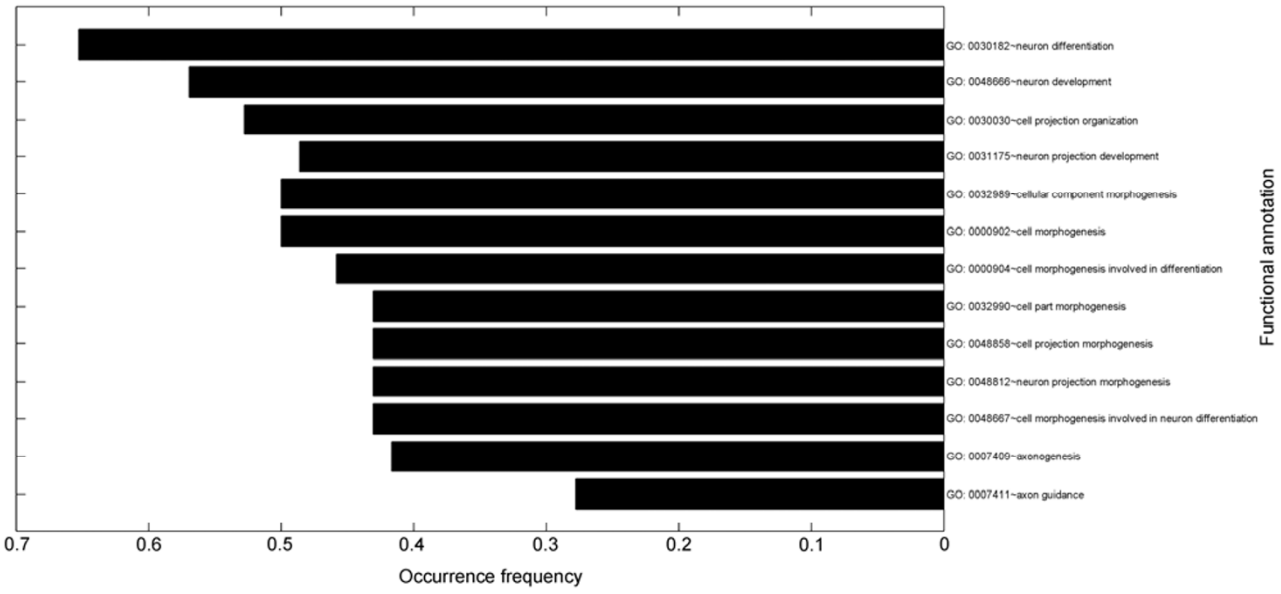

Figure 4 Distribution of functional annotation terms in the most significant cluster.

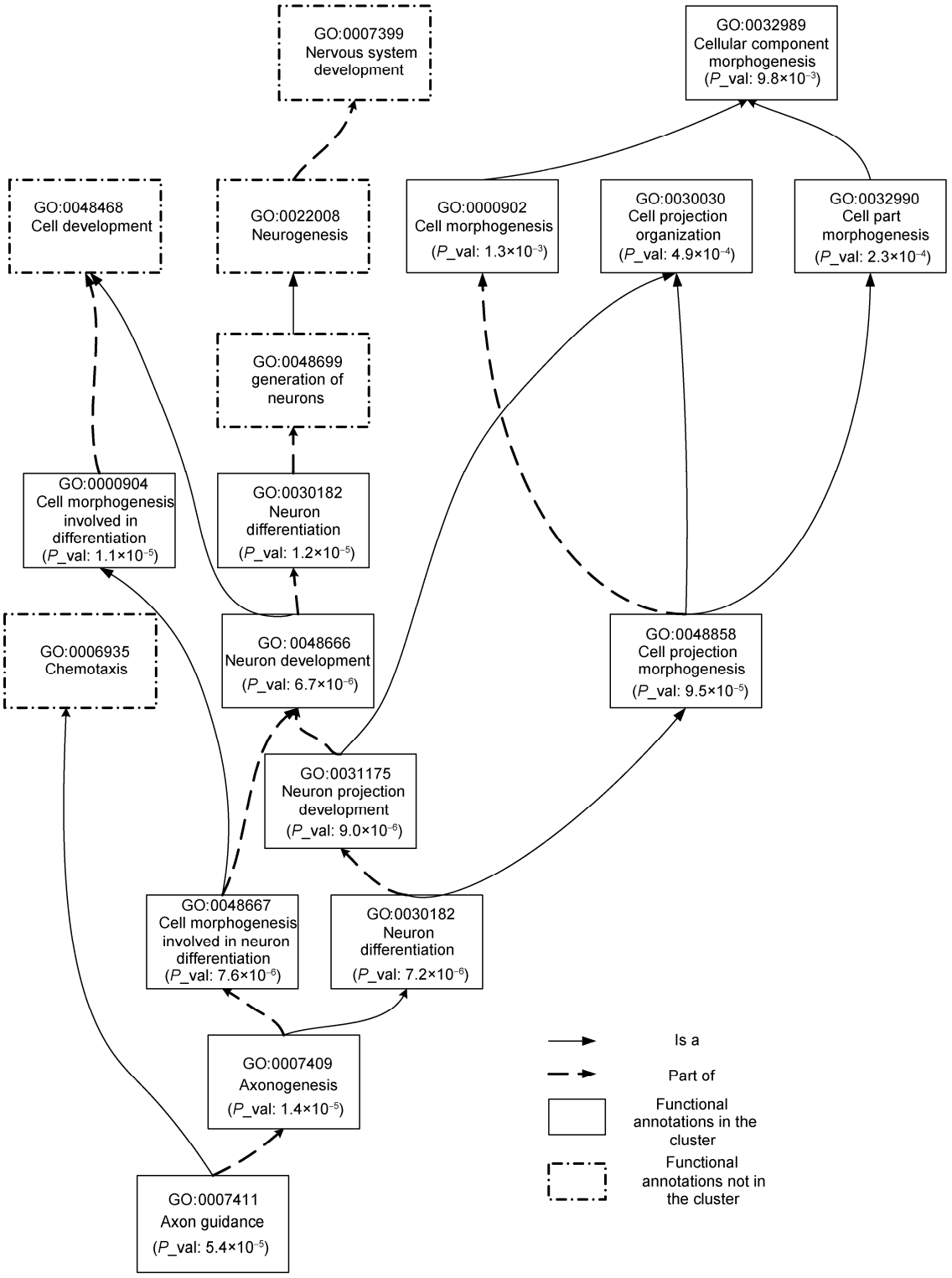

Figure 5 Topology map of the GO annotation terms in the most significant cluster.

The strategy of combining random forests and enrichment analysis takes into account the effect of SNP interactions, and can detect disease-associated SNPs and the corresponding biological mechanisms more sensitively. Therefore, this strategy addresses some of the drawbacks of the current analytical methods that are used in GWA studies. Thus, using random forests to generate candidate SNPs is an efficient alternative to the common enrichment analysis approaches. Furthermore, the proposed method can be easily generalized and applied to GWA studies of other com-

plex diseases. However, it should be noted that compared with the traditional univariate methods, the importance of the features from random forests cannot be used to directly reflect the statistical significance of the detected SNPs. Because both random forests and enrichment analysis are exploratory methods, further biological experiments are required to confirm the findings reported in this study.

This work was supported by the National Natural Science Foundation of China (Nos. 2100230024 and 2100230023).

1 Pandey P, Singh M, Gambhir I. Alzheimer's disease: A Threat to mankind. J Stress Physiol Biochem, 2011, 7: 15–30
2 Ferri C P, Prince M, Brayne C, *et al*. Global prevalence of dementia: a Delphi consensus study. Lancet, 2005, 366: 2112–2117
3 Gatz M, Reynolds C A, Fratiglioni L, *et al*. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry, 2006, 63: 168–174
4 Holscher C. Diabetes as a risk factor for Alzheimer's disease: insulin signalling impairment in the brain as an alternative model of Alzheimer's disease. Biochem Soc Trans, 2011, 39: 891–897
5 Cai J, Yin D. Research progress on important genes and functional proteins related to Alzheimer's disease. Chin J Neuroimmunol Neurol, 2006, 13: 120–123
6 Saunders A M, Strittmatter W J, Schmechel D, *et al*. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology, 1993, 43: 1467–1472
7 Liu Q, Wu W, Li R, *et al*. Advance in research of apolipoprotein E and Alzheimer's disease. Process Chem, 2008, 19: 2006–2011
8 Zhuang Y, Chen J. Research progress on causes and mechanism of Alzheimer's disease. J Jilin Med College, 2008, 29: 1–2
9 Reiman E M, Webster J A, Myers A J, *et al*. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. Neuron, 2007, 54: 713–720
10 Tang L, Lv Z, Yang Z, *et al*. Association between cholesterol 24S-hydroxylase gene polymorphism and late onset Alzheimer disease. Chin J Geriatr, 2007, 26: 13–15
11 Tan L, Liu R, Lei S, *et al*. A genome-wide association analysis implicates SOX6 as a candidate gene for wrist bone mass. Sci China Life Sci, 2010, 53: 1065–1072
12 Wang M, Chen X, Zhang M, *et al*. Detecting significant single-nucleotide polymorphisms in a rheumatoid arthritis study using random forests. BMC Proc, 2009, 3: S69
13 Wang M, Zhang M, Chen X, *et al*. Detecting genes and gene-gene interactions for age-related macular degeneration with a forest-based approach. Stat Biopharm Res, 2009, 1: 424–430
14 Chen X, Liu C T, Zhang M, *et al*. A forest-based approach to identifying gene and gene-gene interactions. Proc Natl Acad Sci USA, 2007, 104: 19199-19203
15 Wang M, Chen X, Zhang H. Maximal conditional chi-square importance in random forests. Bioinformatics, 2010, 26: 831–837
16 Bertram L, Lill C M, Tanzi R E. The genetics of Alzheimer disease: back to the future. Neuron, 2010, 68: 270–281
17 Harold D, Abraham R, Hollingworth P, *et al*. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet, 2009, 41: 1088–1093
18 Satake W, Nakabayashi Y, Mizuta I, *et al*. Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nat Genet, 2009, 41: 1303–1307
19 Han J, Zhang X. Current status of genome-wide association study. Hereditas, 2011, 33: 25–35
20 Birnbaum S, Ludwig K U, Reutter H, *et al*. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. Nat Genet, 2009, 41: 473–477
21 Lunetta K L, Hayward L B, Segal J, *et al*. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet, 2004, 5: 32
22 Dennis G Jr., Sherman B T, Hosack D A, *et al*. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol, 2003, 4: P3
23 Huang da W, Sherman B T, Lempicki R A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc, 2009, 4: 44–57
24 Breiman L. Random forests. Mach learn, 2001, 45: 5–32
25 Trevor H, Robert T, Jerome F. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag, 2001. 371–406
26 Zhang H, Wang M, Chen X. Willows: a memory efficient tree and forest construction package. BMC Bioinformatics, 2009, 10: 130
27 Huang da W, Sherman B T, Lempicki R A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res, 2009, 37: 1–13
28 Guo H, Zhu Y P, Li D, *et al*. Identification, modeling and simulation of key pathways underlying certain cancers. Hereditas, 2011, 33: 809–819
29 Liu M, Wang M, Ding W, *et al*. Gene function enrichment analysis of microarray data. J Biomed Engineer, 2010, 27: 1166–1168
30 Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. Bioinformatics, 2004, 20: 578–580
31 Rivals I, Personnaz L, Taing L, *et al*. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics, 2007, 23: 401–407
32 Hosack D A, Dennis G Jr., Sherman B T, *et al*. Identifying biological themes within lists of genes with EASE. Genome Biol, 2003, 4: R70
33 Genuer R, Poggi J M, Tuleau C. Random Forests: some methodological insights. Arxiv preprint arXiv:08113619, 2008
34 Bertram L, Tanzi R E. Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. Nat Rev Neurosci, 2008, 9: 768–778
35 Zhong X L, Yu J T, Hou G Y, *et al*. Common variant in GRB2 is associated with late-onset Alzheimer's disease in Han Chinese. Clin Chim Acta, 2010, 412: 446–449
36 Lucatelli J F, Barros A C, Silva V K, *et al*. Genetic influences on Alzheimer's disease: evidence of interactions between the genes *APOE*, *APOC1* and *ACE* in a sample population from the south of Brazil. Neurochem Res, 2011, 1–7
37 Bertram L, McQueen M B, Mullin K, *et al*. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet, 2007, 39: 17–23
38 Zhang G, Song H, Chen Z. Molecular mechanism of mRNA alternative splicing. Acta Genet Sin, 2004, 31: 102–107
39 Tollervey J R, Wang Z, Hortobagyi T, *et al*. Analysis of alternative splicing associated with aging and neurodegeneration in the human brain. Genome Res, 2011, 21: 1572–1582
40 Mukai F, Ishiguro K, Sano Y, *et al*. Alternative splicing isoform of tau protein kinase I/glycogen synthase kinase 3beta. J Neurochem, 2002, 81: 1073–1083
41 Li M, Chang X, Tao X. Senile dementia of the Alzheimer type and the abnormal modification of tau protein. J Shantou Univ Med College, 2000, 13: 73–75
42 Tojima T, Ito E. Signal transduction cascades underlying de novo protein synthesis required for neuronal morphogenesis in differentiating neurons. Prog Neurobiol, 2004, 72: 183–193
43 Perez R G, Zheng H, Van der Ploeg L H, *et al*. The beta-amyloid precursor protein of Alzheimer's disease enhances neuron viability and modulates neuronal polarity. J Neurosci, 1997, 17: 9407–9414